SELECTING TAXA TO SAVE OR SEQUENCE: DESIRABLE CRITERIA AND A GREEDY SOLUTION

Magnus BORDEWICH, Allen G. RODRIGO and Charles SEMPLE

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2007/6 DECEMBER 2007

SELECTING TAXA TO SAVE OR SEQUENCE: DESIRABLE CRITERIA AND A GREEDY SOLUTION

MAGNUS BORDEWICH, ALLEN G. RODRIGO, AND CHARLES SEMPLE

ABSTRACT. Three desirable properties for any method of selecting a subset of evolutionary units (EUs) for conservation or for genomic sequencing are discussed. These properties are: *spread, stability,* and *applicability.* We are motivated by a practical case in which the maximisation of phylogenetic diversity (PD), which has been suggested as a suitable method, appears to lead to counter-intuitive collections of EUs and does not meet these three criteria. We define a simple greedy algorithm (GREEDYMMD) as a close approximation to choosing the subset that maximises the minimum pairwise distance between EUs. This method of selection satisfies our three criteria and may be a useful alternative to PD in certain real world situations. We also show that if distances between EUs are ultrametric, then GREEDYMMD delivers an *optimal* subset of EUs that maximises both the minimum pairwise distance and the PD. Finally, since GREEDYMMD works with distances and does not require a tree, it is readily applicable to many datasets.

1. INTRODUCTION

Quantitative methods based on biodiversity have received much attention recently in conservation biology. These methods are used for determining which collection of evolutionary units (EUs, including species or other higher-level taxa)

Date: 1 November 2007.

Key words and phrases. biodiversity conservation, phylogenetic diversity, greedy algorithm.

should be conserved. Maximising phylogenetic diversity (PD) has emerged as a leading criteria in this regard. (For a set of EUs and a given phylogeny, the PD of the set is defined as the total length of the minimal phylogenetic subtree that connects the EUs in the set.) In its most direct application to conservation (Faith, 1992), we want to select a k-element subset of EUs that maximises PD over all k-element subsets. The use of PD has also been advocated when decisions need to be made about which genomes should be sequenced (Pardi and Goldman, 2005).

However, there are instances when maximising PD appears to be the wrong criterion to use when selecting a set of EUs for conservation, or deciding which genomes should be sequenced. Perhaps the best example is the iconic small-subunit ribosomal RNA tree of the bacteria, archaea, and eucarya first described in Woese (1987), and reproduced in Fig. 1. Suppose that the aim is to select three of these EUs for sequencing. A subset of three EUs chosen to maximise PD will include two eukaryotes and one bacterium, as shown in red in Fig. 1A. Most people would agree that this seems wrong; after all, the second eukaryote may add comparatively little information. In contrast, selecting a bacterium, an archaeon, and an eukaryote (as shown in red in Fig. 1B) is a choice that few would disagree with; it seems intuitively better. If one were asked to explain why this is more intuitive, one might reply that the chosen EUs are more "spread out" on the phylogenetic tree, and are likely to contain less redundant genetic information (measured, say, by the number of genes in common).

Since PD does not entirely capture what we want from a measure of diversity, we propose an alternative. The first step is to identify exactly what it is we are after. We do not simply require a measure of diversity, but also a method of selecting a

subset of EUs that will score well under this measure. In addition, we would like the selected set to have the following properties:

(1) **Spread.** The selected EUs are in some sense spread evenly across the phylogeny.
(2) **Stability.** The core of the solution set should be stable as the set size fluctuates.
(3) **Applicability.** The method of selection should be able to be applied to as general an input as possible.

We expand on these properties and why they are desirable below.

Spread. We have discussed one example of where PD does not capture the intuitive solution because its optimal solution sets are not "spread out" on the phylogeny. In fact, the idea of choosing a set of EUs that have this property of being "spread out" along a tree has been formalised by Holland (2001). In particular, to obtain an intuitively appropriate solution one needs to choose the set of EUs that maximises the minimum phylogenetic distance between any pair of EUs in the set. Defined formally in Section 2, we refer to this criterion as MMD (Maximise Minimum-Distance). If MMD is applied to choose three of the EUs in Fig. 1, then a desired set of one bacterium, one archaeon and one eukaryote will be obtained.

Stability. The amount of money available for conservation projects fluctuates with the political climate and with other social and economic factors. The set of species targeted for conservation or set of genomes chosen to sequence needs to be stable as budgets vary. Suppose that you are told that there is sufficient money to conserve k species, or sequence k genomes. Being a good phylogeneticist,

you apply a diversity-based method to choose your k-element subset of EUs. Your project gets underway, but after a year the budget available is adjusted up or down. Clearly, it is extremely desirable that the set of EUs selected under the new budget is closely related to the original set of EUs chosen. If the number of EUs can be increased, you would like the k EUs previously chosen, and partially worked on, to be guaranteed to be in the larger set of EUs chosen by the algorithm you have used for selection. If the number of EUs is to be reduced, you would like the new solution set to be a subset of the EUs you have already started conserving or sequencing. This notion of stability is captured by a *greedy algorithm*: selecting the pair of EUs that maximises (pairwise) a measure of diversity, and then iteratively selecting an EU to add to the set that gives the best measure amongst all single EU increments to the existing set, until the desired number of EUs is reached. Therefore, any increase in the size of the chosen set is made by simply adding more EUs to the existing set, whilst any decrease may be made by removing the EUs added last. Thus we see that a greedy selection criteria exactly captures the idea of stability that we are after for our method of selection.

Applicability. We would like to be able to apply the selection method to many varied data sets. It is not always the case that an accurate tree with known branch lengths will be available. However, as a minimum, it seems reasonable to expect that a matrix of pairwise distances between taxa will be available, or at least readily estimated from sequences (*e.g.* using Hamming distance). Therefore we would like our method and measure of diversity to apply in this situation, without having to go to the computational expense and loss of information involved in estimating a phylogeny first.

We now consider what method of selecting EUs would satisfy these three criteria. As observed above, the measure MMD seems to capture the notion of spread that we require, whereas PD does not. In addition, since MMD depends only on the pairwise distances between EUs, it can be applied to a distance matrix without reference to any underlying tree. Thus MMD satisfies the applicability criteria, while again PD does not. For stability, the situation is reversed. An optimal set under the PD measure is generated by a greedy algorithm (Pardi and Goldman, 2005; Steel, 2005), while under MMD being greedy does not always return the optimal solution (see Example 3.4). Hence MMD will not necessarily have the stability property, whereas PD does. Indeed, after an increase of one EU, the revised optimal subset of EUs under MMD could exclude some of the original EUs you had selected (see Example 3.1). Thus neither MMD nor PD satisfies all three criteria. In order to satisfy all three, we propose a greedy approach to MMD. This method, called GREEDYMMD, gains stability and still maintains a good, though not necessarily optimal, solution to MMD, thus retaining spread and applicability.

In Section 3, we define GREEDYMMD and prove a sharp bound to how accurately it performs as an approximation to MMD. To be precise, we show that GREEDYMMD gives a 2-approximation to the optimal MMD solution provided the pairwise distances satisfy the triangle inequality (Theorem 3.3). Note that this includes the case that the distances fit a tree metric. Furthermore, if the distance matrix satisfies the triangle inequality but is not necessarily a tree metric, then it is NP-hard to compute an optimal set under MMD. Moreover, no polynomial-time algorithm can consistently return a better approximation than that given by GREEDYMMD, unless P=NP (see Section 3). In addition, we show that if the

pairwise distances induce an ultrametric, then GREEDYMMD returns an *optimal* set of EUs under MMD and, moreover, this set also maximises PD (see Section 4).

One possible criticism of MMD as a measure of diversity is that it depends only on the closest pair of EUs. Thus one can imagine a situation in which there are t well separated clades, each consisting of a large number of closely related EUs. Any set of $t+1$ EUs will have to contain two EUs from one of the clades. Hence, under MMD, the measure for a set consisting of $t+1$ EUs from a single clade will be similar to the measure for any other set of $t+1$ EUs. This example is another illustration of why being greedy is good. The greedy solution for k EUs will always contain the solution for j EUs for every $j < k$. Thus in a situation similar to that outlined here, our GREEDYMMD solution for $t+1$ EUs will contain the solution for t EUs, *i.e.* at least one from every clade.

In consequence of the above analysis, we conclude that the selection method GREEDYMMD satisfies the three criteria of spread, stability and applicability, and advocate it as a useful and practical alternative to PD in certain real world situations.

The paper is organised as follows. The next section contains some preliminaries and a formal definition of MMD. In Section 3, we present GREEDYMMD and show that it provides a 2-approximation to MMD if the pairwise distances satisfy the triangle inequality, while Section 4 considers GREEDYMMD applied to a set of pairwise distances that induce an ultrametric. In Section 5, we consider taxonomic distinctness, a measure for diversity with similarities to MMD. The last section summarises the paper with a brief discussion.

2. Definitions and Problem Specification

In this section, we detail some definitions that are used throughout the paper and formally describe MMD.

An (unrooted) *phylogenetic X-tree* is a tree with no degree-two vertices and whose leaf set is X. A *rooted phylogenetic X-tree* is a rooted tree with no degree-two vertices except the root which may have degree two and whose leaf set is X. For the purposes of this paper, we will assume that all the edges of a (rooted or unrooted) phylogenetic tree are assigned non-negative real-valued lengths.

For a set X, a *distance* on X is a function δ that assigns a non-negative real value to each ordered pair in $X \times X$ such that, for all $x, y \in X$, we have $\delta(x, x) = 0$ and $\delta(x, y) = \delta(y, x)$. A distance is said to satisfy the *triangle inequality* if, for all $x, y, z \in X$,

$$\delta(x, z) \leq \delta(x, y) + \delta(y, z).$$

A natural way to obtain a distance on X is from a phylogenetic X-tree. In particular, a distance $\delta_{\mathcal{T}}$ on X can be obtained by setting $\delta_{\mathcal{T}}(x, y)$ to be the sum of the edge-lengths on the (unique) path from x to y for all $x, y \in X$. Distances that can be realised via a phylogenetic tree in this way are known as *tree metrics*. Such metrics satisfy the triangle inequality.

A tree metric $\delta_{\mathcal{T}}$ on X is an *ultrametric* if it can be realised by a rooted phylogenetic X-tree $\mathcal{T}$ such that, for all $x, y \in X$,

$$\delta_{\mathcal{T}}(\rho, x) = \delta_{\mathcal{T}}(\rho, y), \tag{1}$$

where ρ is the root of $\mathcal{T}$. The equality in (1) means that all leaves are equidistant from the root. Equivalently, for an arbitrary distance δ on X, δ is an ultrametric precisely if, for every three distinct elements $x, y, z \in X$,

$$\delta(x, y) \leq \max\{\delta(x, z), \delta(y, z)\}.$$

For this equivalence, see Semple and Steel (2003).

Let δ be a distance on X. For a subset S of X, we define the *minimum distance* of S to be

$$MD(S) = \min\{\delta(x, y) : x, y \in S\}.$$

For a given positive integer k, the problem MMD is to find a subset S of X that maximises $MD(S)$ amongst all subsets of X of size k. Intuitively, such a set S corresponds to selecting a k-element subset of X in which each pair of elements is as 'far apart' as possible under δ. Formally, the problem is defined as follows:

Problem: MMD

Instance: A set X, a distance δ on X, and a positive integer k.

Question: Find a subset S of X of size k that maximises $MD(S)$.

As noted in the introduction, whereas it is usual to think of phylogenetic diversity as a criterion that requires a phylogenetic tree as a basis for implementation, MMD is more general. In particular, the criterion MMD can be applied to any distance measure.

3. A Greedy 2-Approximation Algorithm for MMD

In this section, we analyse the simple greedy approach for selecting a subset of EUs under MMD with distance δ. If δ is a tree metric, then, as observed in Spillner et al. (2007), one can obtain an optimal solution for MMD in polynomial time using the techniques of Chandrasekaran and Daughety (1981). Nevertheless, because of the desirable property of stability, we are interested in greedy solutions in their own right, and so wish to understand their relative performance. Example 3.1 illustrates the potential problem of using the true optimal solution.

Example 3.1. Consider the phylogenetic tree shown in Fig. 2. Under MMD, it is easily checked that $S_5 = \{x_1, x_3, x_5, x_7, x_9\}$ is the unique optimal set of five taxa ($MD(S_5) = 7$), while $S_6 = \{x_1, x_2, x_4, x_6, x_8, x_9\}$ is the unique optimal set of six taxa ($MD(S_6) = 6$). Thus, in this instance, increasing resources to enable an additional taxa to be conserved would see three of the currently selected taxa dropped from the optimal set. This example can be extended in the obvious way so that, for an arbitrary k, the unique optimal set of size $k+1$ intersects in only two taxa with the unique optimal set of size k.

We analyse the following greedy algorithm.

GreedyMMD(δ, k)

Step 1 Let S be the empty set.

Step 2 Select the two most distant EUs and add to S.

(That is, select two elements x and y that maximises $\delta(x, y)$.)

Step 3 Set counter $c = 2$.

Step 4 If $c = k$, STOP; otherwise, select an EU from those not already included in S so that the minimum distance between that EU and those in S is maximum amongst all remaining EUs not in S.

(That is, select $z \in X - S$ that maximises $\min_{y \in S} \delta(z, y)$.)

Step 5 Set $c = c + 1$ and return to Step 4.

The next theorem shows that provided δ satisfies the triangular inequality, then GREEDYMMD is a 2-approximation algorithm to MMD. This means that if S is the solution returned by GREEDYMMD and Y_{opt} is an optimal solution, then $2MD(S) \geq MD(Y_{\text{opt}})$, that is, $MD(S)$ is at least half $MD(Y_{\text{opt}})$. It is shown in Ravi et al. (1994) that, assuming δ is only guaranteed to satisfy the triangle inequality, for any $\epsilon > 0$, no polynomial-time algorithm can return a $(2 - \epsilon)$-approximation to MMD unless P=NP. In particular, in this case the problem MMD is NP-hard. Hence, in this setting, GREEDYMMD gives the best possible approximation. This theorem has previously been observed in the context of operations research, for example, see Tamir (1991) and Ravi et al. (1994). However, we include our proof here as we will require the preliminary lemma again in the next section and the proof is written in the language of phylogenetics.

Lemma 3.2. *Let δ be a distance on X and let k be an integer greater than two. Let S_{k-1} be the $(k-1)$-element set that is constructed at the completion of the second-to-last iteration of* GREEDYMMD(δ, k). *Then, for any element $x \in X - S_{k-1}$, we have $MD(S_{k-1} \cup \{x\}) = \delta(x, s)$ for some $s \in S_{k-1}$.*

Proof. For $2 \leq i \leq k-1$, let $S_i = \{s_1, s_2, \ldots, s_i\}$ denote the i-element subset of S_{k-1} that is sequentially constructed by GREEDYMMD(δ, k). If $MD(S_{k-1} \cup \{x\}) \neq$

$\delta(x, s)$ for any $s \in S_{k-1}$, then, for some distinct $s_i, s_j \in S_{k-1}$ with $i < j$, we have $MD(S_{k-1} \cup \{x\}) = \delta(s_i, s_j) < \delta(x, s)$. If there is more than one pair s_i, s_j, choose a pair with minimal j, and so $MD(S_{j-1}) > \delta(s_i, s_j)$. But then, for all $s \in S_{j-1}$,

$$MD(S_{j-1} \cup \{x\}) \geq \min\{MD(S_{j-1}), \min_{s \in S_{j-1}} \delta(x, s)\} > \delta(s_i, s_j) \geq MD(S_j),$$

contradicting the way in which GREEDYMMD selects elements. The lemma now follows. □

Theorem 3.3. *Let δ be a distance on X, and suppose that δ satisfies the triangle inequality. Let k be an integer greater than one and let S_k be set returned by* GREEDYMMD(δ, k). *Then $MD(S_k)$ is a 2-approximation to $MD(Y_{\text{opt}})$, where Y_{opt} is an optimal solution of size k to* MMD.

Proof. For all $2 \leq i \leq k$, let $S_i = \{s_1, s_2, \ldots, s_i\}$ denote the i-element subset of S_k that is sequentially constructed by GREEDYMMD(δ, k). By Lemma 3.2, $MD(S_k) = \delta(s_k, s_i)$ for some $i \in \{1, 2, \ldots, k-1\}$.

Let $Y_{\text{opt}} = \{y_1, y_2, \ldots, y_k\}$ be an optimal solution of size k to MMD. Amongst the elements in $Y_{\text{opt}} - S_{k-1}$, let y be an element such that

$$MD(S_{k-1} \cup \{y\}) = \max\{MD(S_{k-1} \cup \{y'\}) : y' \in Y_{\text{opt}} - S_{k-1}\}.$$

By Lemma 3.2, $MD(S_{k-1} \cup \{y\}) = \delta(y, s_{i'})$ for some $i' \in \{1, 2, \ldots, k-1\}$. Because of the selection criteria of GREEDYMMD, $MD(S_k) \geq MD(S_{k-1} \cup \{y\})$, and so $\delta(s_k, s_i) \geq \delta(y, s_{i'})$.

Now assign each element of Y_{opt} to the element in S_{k-1} that it is closest to under δ. Since $|Y_{\text{opt}}| > |S_{k-1}|$, it follows by the pigeon-hole principle that two distinct

elements $y_r, y_s \in Y_{\text{opt}} - S_{k-1}$ are assigned to the same element, s say, in S_{k-1}. By the choice of y above,

$$\delta(y_r, s) \leq \delta(y, s_{i'}) \text{ and } \delta(y_s, s) \leq \delta(y, s_{i'}).$$

Then, as δ satisfies the triangle inequality and it is symmetric,

$$\begin{aligned} MD(Y_{\text{opt}}) &\leq \delta(y_r, y_s) \\ &\leq \delta(y_r, s) + \delta(s, y_s) = \delta(y_r, s) + \delta(y_s, s) \\ &\leq 2\delta(y, s_{i'}) \\ &\leq 2\delta(s_k, s_i) \\ &= 2MD(S_k) \\ &\leq 2MD(Y_{\text{opt}}). \end{aligned}$$

That is,

$$MD(Y_{\text{opt}})/2 \leq MD(S_k) \leq MD(Y_{\text{opt}}).$$

Hence $MD(S_k)$ is a 2-approximation to $MD(Y_{\text{opt}})$. □

It is possible that for a given set of EUs, the k-element subset returned by GREEDYMMD and that which optimises MMD will exhibit the same minimum distance or, in fact, be the same subsets. However, we reiterate that the greedy algorithm cannot be guaranteed to work any better than as a 2-approximation algorithm even if δ is a tree metric as we show in the following example. We say, therefore, that Theorem 3.3 is sharp.

Example 3.4. Consider the phylogenetic X-tree $\mathcal{T}$ shown in Fig. 3, where the length of the edge incident with z is arbitrarily small. Suppose that GREEDYMMD

is applied to the distance on X induced by the path lengths in $\mathcal{T}$, where k is chosen so that $t+2 \leq k \leq 2t$. Let S be the subset of X returned by this application. Without loss of generality, we may assume that the first t elements of X selected by GREEDYMMD are $x_1, x_3, \ldots, x_{2t-1}$. The $(t+1)$-th element selected by GREEDYMMD is z, while the remaining elements are any $k-(t+1)$ elements in $\{y_1, y_2, \ldots, y_t\}$. A simple check shows that $MD(S) = 3+\epsilon$. However, it is easily seen that

$$Y_{\text{opt}} = \{x_1, x_3, \ldots, x_{2t-1}, y_1, y_2, \ldots, y_{k-t}\}$$

is an optimal solution and $MD(Y_{\text{opt}}) = 6$. Thus the approximation ratio in this case can be made arbitrarily close to 2 by an appropriate choice of ϵ.

4. ULTRAMETRIC DISTANCES

In the previous section, we described how well GREEDYMMD performed as an approximation to MMD provided the distance measure satisfied the triangle inequality. In this section, we show that there are conditions under which GREEDYMMD will always return a subset of EUs that is optimal under MMD. In particular, the next theorem guarantees that if δ is an ultrametric, then the solution set S returned by GREEDYMMD(δ, k) is an optimal solution of size k to MMD. Moreover, if $\mathcal{T}$ is a rooted phylogenetic tree that realises δ, then the PD score of S, denoted $PD(S)$, is equal to the maximum PD score over all subsets of size k. We denote this last score by PD_k. Note that, in this setting, $PD(S)$ is the total length of the subtree connecting the elements of S and the root of the phylogeny.

Theorem 4.1. *Let δ be an ultrametric on X, and let k be an integer greater than one. Let S_k be the set returned by* GREEDYMMD(δ, k) *and let Y_{opt} be an optimal*

solution of size k to MMD. *Then*

$$MD(S_k) = MD(Y_{\text{opt}}).$$

Moreover, if $\mathcal{T}$ is a rooted phylogenetic X-tree that realises δ, then $PD(S_k) = PD_k$ on $\mathcal{T}$.

Proof. The first part of the proof proceeds in a similar way to that used for proving Theorem 3.3. Nevertheless, for clarity, we include the proof in full.

For $2 \leq i \leq k$, let $S_i = \{s_1, s_2, \ldots, s_i\}$ denote the i-element subset of S_k that is sequentially constructed by GREEDYMMD(δ, k). By Lemma 3.2, $MD(S_k) = \delta(s_k, s_i)$ for some $i \in \{1, 2, \ldots, k-1\}$. Let $Y_{\text{opt}} = \{y_1, y_2, \ldots, y_k\}$ be an optimal solution of size k to MMD. Amongst all elements in $Y_{\text{opt}} - S_{k-1}$, let y be an element that maximises $MD(S_{k-1} \cup \{y\})$. By Lemma 3.2, $MD(S_{k-1} \cup \{y\}) = \delta(y, s_{i'})$ for some $i' \in \{1, 2, \ldots, k-1\}$. Now $\delta(s_k, s_i) \geq \delta(y, s_{i'})$ as $MD(S_k) \geq MD(S_{k-1} \cup \{y\})$. Assign each element of Y_{opt} to the element of S_{k-1} that it is closest to under δ. Since $|Y_{\text{opt}}| > |S_{k-1}|$, there are two distinct elements y_r and y_s in $Y_{\text{opt}} - S_{k-1}$ assigned to the same element s in S_{k-1}. By choice of y above,

$$\delta(y_r, s) \leq \delta(y, s_{i'}) \text{ and } \delta(y_s, s) \leq \delta(y, s_{i'}).$$

Thus, since δ is an ultrametric, we have

$$\begin{aligned} MD(Y_{\text{opt}}) &\leq \delta(y_r, y_s) \\ &\leq \max\{\delta(y_r, s), \delta(y_s, s)\} \\ &\leq \delta(y, s_{i'}) \\ &\leq \delta(s_k, s_i) \\ &= MD(S_k) \\ &\leq MD(Y_{\text{opt}}). \end{aligned}$$

In other words, equality holds throughout and so $MD(S_k) = MD(Y_{\text{opt}})$. This completes the first part of the theorem.

To show that $PD(S_k) = PD_k$ on $\mathcal{T}$, we use induction on k. Clearly, the result holds if $k = 2$. So assume that the result holds whenever the set return by GREEDYMMD has size at most $k-1$, where $k \geq 3$.

Suppose that $S_k = \{s_1, s_2, \ldots, s_k\}$ is returned by GREEDYMMD. By Lemma 3.2, $MD(S_k) = \delta(s_k, s_i)$ for some $i < k$. Let u be the most recent common ancestor of s_k and s_i in $\mathcal{T}$. Then, as δ is an ultrametric,

$$MD(S_k) = \delta(s_i, s_k) = \delta(s_i, u) + \delta(u, s_k) = 2\delta(u, s_k).$$

Thus, in determining S_k, GREEDYMMD finds the element s in $X - S_{k-1}$ that maximises the sum of the edge lengths from s to the minimal subtree of $\mathcal{T}$ that connects the elements in S_{k-1}. Indeed, this is exactly how the greedy algorithm works in selecting elements for optimising phylogenetic diversity. By the induction

assumption, $PD(S_{k-1}) = PD_{k-1}$, and so $PD(S_k) = PD_k$. This completes the proof of the theorem. □

5. Taxonomic Distinctness

A measure closely related to MMD that is widely used in conservation biology, though in slightly different circumstances, is *taxonomic distinctness* (TD), see Clarke and Warwick (1998). Typically, TD is used for comparing the biodiversity of different areas. Each area is visited and the set of taxa observed within the area is recorded. The TD of the area is (effectively) taken to be the average pairwise distance between taxa. This naturally leads to the idea in our setting of selecting a set of EUs that maximises the *average* distance between EUs (MAD). At first sight, this appears similar to maximising the minimum pairwise distance as in MMD. However, for many instances in which there is dominant longest path in the tree, *e.g.* to an out-group of EUs, the optimal and greedy sets chosen under MAD unduly try to balance the number of EUs either side of this path, see Example 5.1. A greedy set under MAD is chosen in the same way as that for MMD except that Step 4 in GreedyMMD is replaced with the following step:

Step 4′ If $c = k$, STOP; otherwise, select an EU from those not already included in S so that the sum of the pairwise distances between that EU and those in S is maximum amongst all remaining EUs not in S.
(That is, select $z \in X - S$ that maximises $\sum_{y \in S} \delta(z, y)$.)

Example 5.1. Consider the ultrametric tree shown in Fig. 4. Under MMD and PD, a greedy (and therefore optimal) solution set will involve an a_i and a spread

of elements from $\{b_1, b_2, \ldots, b_m\}$. Whereas, under MAD, both greedy and optimal solution sets will have half of their elements from $\{a_1, a_2, \ldots, a_n\}$.

Likewise, when three EUs from Fig. 1 are selected to optimise MAD, two eukaryotes and one bacterium are chosen, as in PD. Thus MAD does not capture the notion of spread we desire.

For completeness, we note below the computational similarities and differences between MMD and MAD.

(1) If δ is a tree metric, then, as in the case for MMD, selecting an optimal set of k EUs under MAD can be done in polynomial time (Chandrasekaran and Daughety, 1981).

(2) If δ is a tree metric, then, as far as we are aware, the exact approximation ratio of the greedy algorithm for MAD is unknown. Ravi et al. (1994) have shown that it is no worse than a 4-approximation; simple examples show that it is no better than a 4/3-approximation. In comparison, the greedy algorithm GREEDYMMD returns a 2-approximation to MMD and this is sharp.

(3) If δ satisfies the triangle inequality, Ravi et al. (1994) show that the greedy algorithm for MAD is no worse than a 4-approximation and no better than a 2-approximation algorithm. Again, GREEDYMMD returns a 2-approximation to MMD, and this is sharp.

6. Discussion

In this paper, we show that a subset of EUs that maximises phylogenetic diversity may result in a counter-intuitive collection of species earmarked for conservation or as genomic sequencing targets. We argue that as a plausible alternative to phylogenetic diversity, an appropriate criterion is to choose the subset of EUs where the minimum distance between any pair of EUs is maximum, amongst all possible subsets of the same size. This criterion has the virtue of choosing EUs that are "spread out" across the phylogenetic tree.

However, as others have shown (e.g. Moulton et al., 2007), choosing the subset that maximises the minimum pairwise distance between EUs cannot be achieved by applying a greedy algorithm whereby one "builds up" a subset of k EUs by successively adding to optimal subsets of $2, 3, \ldots, k-1$ EUs. Nonetheless, we think that there is real value in being greedy. When one has the opportunity to sequence yet another genome or save yet another species, one would want all the genomes that have already been sequenced and all the species that have already been saved to be part of the larger optimal subset. A greedy algorithm guarantees this.

It is of value, then, to consider the extent to which a greedy algorithm approximates an optimal implementation of MMD. We showed that, at worst, the greedy algorithm we defined (GreedyMMD) will choose EUs that are separated on a tree by distances no shorter than half the shortest distance of an optimal subset of EUs under MMD provided the distance satisfies the triangle inequality. This suggests that GreedyMMD will still choose a subset that is reasonably "spread out".

When distances are ultrametric, Theorem 4.1 shows that GREEDYMMD returns a set which is in fact optimal under both MMD and PD. This is significant because, until now, the only way to find a set which maximises PD has been to use a phylogenetic tree. Here, we can apply GREEDYMMD to obtain the set of EUs that maximise PD by working only on the distance matrix. One may question its applicability, given that ultrametric distances are unlikely to apply to real EUs. However, it may be possible that even with approximately ultrametric distances GREEDYMMD produces a set that is close to optimal under PD.

Lastly, GREEDYMMD can be used on a pairwise distance matrix that does not necessarily induce a tree metric. Thus a large set of sequences could be analysed using this method based upon a simple distance measure such as Hamming distance, without expending the computational time required to reconstruct an accurate phylogeny. Moreover, if our data satisfies the triangle inequality, we have the performance guarantee of Theorem 3.3.

ACKNOWLEDGEMENTS

The first author was supported by the EPSRC, the second was supported by the Allan Wilson Centre for Molecular Ecology and Evolution, and the third author was supported by the New Zealand Marsden Fund. We thank Geoff Nicholls and Mike Steel for useful discussions. This project began at the Isaac Newton Institute for the Mathematical Sciences, under the auspices of the Phylogenetics Programme.

References

Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. 93:9188–9193.

Chandrasekaran, R. and A. Daughety. 1981. Location on tree networks: p-centre and n-dispersion problems. Math. Oper. Res. 6:50–57.

Clarke, K. R. and R. M. Warwick. 1998. A taxonomic distinctness index and its statistical properties. The Journal of Applied Ecology 35:523–531.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61:1–10.

Guindon, S. and O. Gascuel. 2003. A simple, fast , and accurate method to estimate large phylogenies by maximum liklihood. Syst. Biol. 52:696–704.

Holland, B. R. 2001. Evolutionary analyses of large data sets: Trees and beyond. Ph.D. thesis Massey University.

Moulton, V., C. Semple, and M. Steel. 2007. Optimizing phylogenetic diversity under constraints. J. Theoret. Biol. 246:186–194.

Pardi, F. and N. Goldman. 2005. Species choice for comparative genomics: being greedy works. PLoS Genetics 1:e71.

Ravi, S. S., D. J. Rosenkrantz, and G. K. Tayi. 1994. Heuristic and special case algorithms for dispersion problems. Operations Research 42:299–310.

Sanderson, M. J., M. J. Donoghue, W. Piel, and T. Eriksson. 1994. Treebase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Amer. Jour. Bot. 81.

Semple, C. and M. A. Steel. 2003. Phylogenetics. Oxford University Press, Oxford.

Spillner, A., B. Nguyen, and V. Moulton. 2007. Computing phylogenetic diversity for split systems, preprint.

Steel, M. 2005. Phylogenetic diversity and the greedy algorithm. Syst. Biol. 54:527–529.

Tamir, A. 1991. Obnoxious facility location on graphs. SIAM J. Disc. Math. 4:550–567.

Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. 41.

FIGURE CAPTIONS

Figure 1. Reproduction of Woese's (Woese, 1987) small-subunit ribosomal RNA tree showing the subtree subtended by three EUs chosen by (A) minimising PD and by (B) maximising the minimum distance. We constructed this tree using small-subunit ribosomal RNA sequences from an alignment by Barns et al. (1996) available in TreeBase (Sanderson et al., 1994). Maximum likelihood trees were constructed with PHYML (Guindon and Gascuel, 2003) using a GTR model of evolution. The three groups on our tree are represented by the following taxa. ARCHAEA: *Methanococcus vannielli* (Methanogen A), *Methanobacterium* (Methanogen B), *Thermococcus* (Extreme thermophile A), *Thermoproteus* (Extreme thermophile B), *Desulfurococcus* (Extreme thermophile C), *Haloferax* (Extreme halophiles); BACTERIA: *Thermotoga*, *Flavobacteria* (Flavobacteria), *Gloeobacter* (Cyanobacteria), *Escherichia coli* (Purple bacteria), *Bacillus* (Gram-positive bacteria), *Thermomicrobium* (Green non-sulphur bacteria); EUKARYOTES: *Vairimorphal* (Microsporidia), *Euglena geniculata* (Flagellates), *Dictyostelium* (Cellular slime molds), *Zea mays* (Plants), *Homo sapiens* (Animals) and *Coprinus* (Fungi).

Figure 2. A phylogeny on which the optimal set of 5 taxa and the optimal set of 6 taxa selected under MMD intersect in only 2 taxa.

Figure 3. A phylogenetic tree demonstrating that GREEDYMMD cannot be guaranteed to be better than a 2-approximation. For $2 < i < t$, the configuration consisting of leaves y_i, x_{2i-1}, and x_{2i} is repeated where indicated by dashed lines.

Figure 4. An ultrametric tree on which MAD disagrees with MMD and PD.

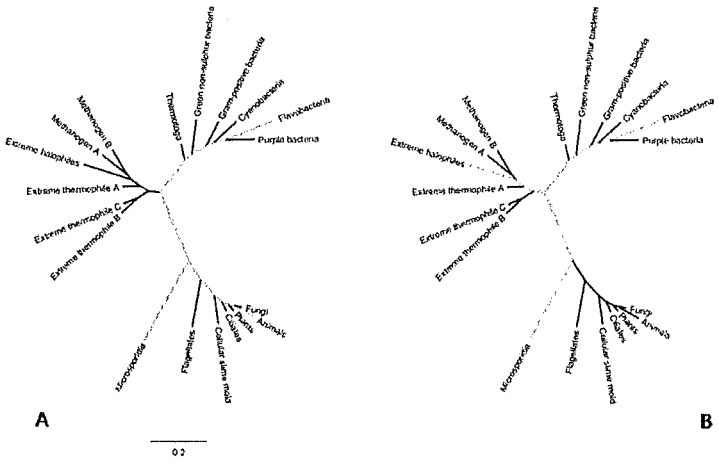
Methanogen B
Methanogen A
Extreme halophiles
Extreme thermophile A
Extreme thermophile C
Extreme thermophile B
Thermotoga
Green non-sulphur bacteria
Gram-positive bacteria
Cytophagales
Flavobacteria
Purple bacteria
Fungi
Animals
Plants
Ciliates
Flagellates
Cellular slime molds
Microsporidia
A
0.2
Methanogen B
Methanogen A
Extreme halophiles
Extreme thermophile A
Extreme thermophile C
Extreme thermophile B
Thermotoga
Green non-sulphur bacteria
Gram-positive bacteria
Cytophagales
Flavobacteria
Purple bacteria
Fungi
Animals
Plants
Ciliates
Flagellates
Cellular slime molds
Microsporidia
B

FIGURE 1

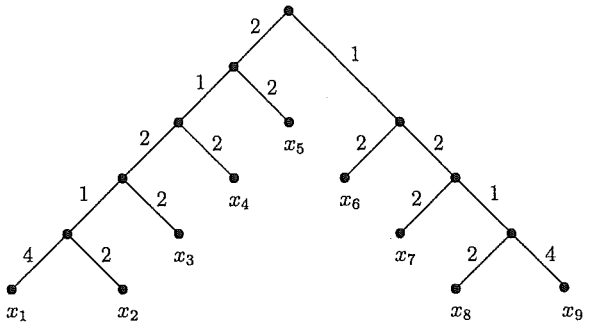

FIGURE 2

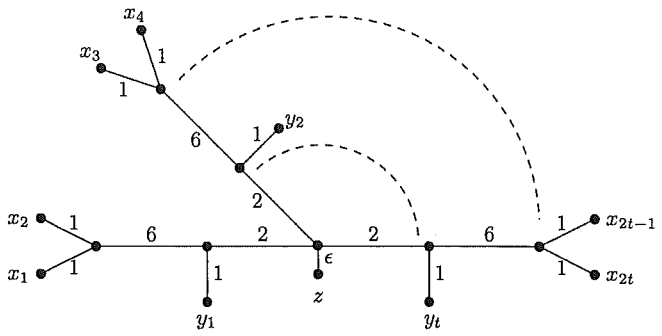

FIGURE 3

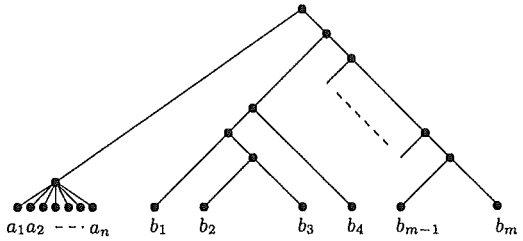

FIGURE 4

MAGNUS BORDEWICH, DEPARTMENT OF COMPUTER SCIENCE, DURHAM UNIVERSITY, DURHAM DH1 3LE, UNITED KINGDOM

E-mail address: m.j.r.bordewich@durham.ac.uk

ALLEN G. RODRIGO, BIOINFORMATICS INSTITUTE AND THE ALLAN WILSON CENTRE FOR MOLECULAR ECOLOGY AND EVOLUTION, UNIVERSITY OF AUCKLAND, AUCKLAND, NEW ZEALAND

E-mail address: a.rodrigo@auckland.ac.nz

CHARLES SEMPLE, BIOMATHEMATICS RESEARCH CENTRE, DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: c.semple@math.canterbury.ac.nz